Analysis of Dust Exposure Impact on Cardiovascular Diseases Risk Prediction in Bangkok, Thailand

Jakkrit Sonsanit, Watcharin Sirigate, Kitsuchart Pasupa
Faculty of Information Technology
King Mongkut’s Institute of Technology Ladkrabang
Bangkok, Thailand
{60070128, 60070158, kitsuchart}@it.kmitl.ac.th

Varut Wiseschinda, Chaiyos Kunanusont
Bangkok Health Research Center
Bangkok Dusit Medical Services PCL
Bangkok, Thailand
{varut.wi, chaiyos.ku}@bdms.co.th

Abstract—Cardiovascular disease (CVD) is a leading cause of death for people around the world. Prediction of CVD risk in advance is one of the most useful and effective tools to prevent and control your risk of developing CVD. Air pollution is a threat to health problems worldwide due to the development of the economy and society. The past studies found that air pollution was one factor that can cause CVD. Additionally, for the Thai population, pollution was one of the factors contributing to premature death. In this work, we aim to predict CVD risk using the patient data set from Bangkok Hospital in Bangkok alone with several algorithms and increase the accuracy using a combination of health and pollution. The prediction based on health data alone with the average AUC scores 0.89±0.03, while the prediction with added air pollution data with the average AUC scores 0.91±0.03 an average 0.02 increase than based on health data results alone. We found that considering pollution data can improve the overall performance of the model to predict CVD risk.

I. INTRODUCTION

Cardiovascular Diseases (CVD) are a group of disorders of the heart and blood vessels, for example, coronary heart disease, cerebrovascular disease, peripheral artery disease, rheumatic heart disease, congenital heart disease and deep vein thrombosis, as well as pulmonary embolism. CVD is a leading cause of death for people around the world. The World Health Organization (WHO) estimated 17.9 million deaths in CVD from 2016, representing 31% of all deaths worldwide [1]. According to a public health statistics report from the Thai Ministry of Public Health [2], for death certificates in 2019, 2 out of 5 deaths were caused by heart disease, and the death toll was 74,445, an increase of 6.82% from 2016. CVD can be prevented and controlled, if the patient is aware of the conditions, that lead to it and the risk of contracting it. Therefore, accurate prediction of CVD disease is essential.

CVD risk can be assessed by observing the initial symptoms, shown by some patients. Additionally, statistical appraisal methods, for example, Framing-ham risk score [3], ASCVD risk score [4], and ThaiCV risk score [5], can predict risk, by processing immutable factors, e.g. age, gender and factors that can be changed, e.g. total cholesterol, blood pressure, smoke status and others, to indicate the likelihood of CVD. However, CVD occurrences in a population, with diverse individuals, still cannot be effectively predicted [6, 7], because each reported prediction model can be applied, with accuracy only applicable to a specific group of individuals, with some distinctive characteristics and a specific type of environment.

Current researches are applying machine learning to CVD prediction, because it promises significantly more effective results, than previous statistical methods. For example, Yang et al. accurately predicted CVD in the eastern population of Zhejian China [6]. Similarly, Alaa et al. predicted CVD in the UK [8].

Air pollution is a major threat to health problems. It causes 4.2 million premature deaths of urban and rural residents worldwide each year, especially people with CVD [9]. Almost all were deaths from stroke, heart disease, chronic obstructive pulmonary disease, lung cancer and acute respiratory infections [10]. The regions most affected by air pollution were the Western Pacific and the developing Southeast Asia. Thailand is one of Southeast Asian regions, that has always faced pollution problems, especially air pollution from fine particles. Thailand set a national standard for level of particulate matter, as small as 2.5 microns (PM2.5) in the air in 2010, but very few Thais were aware of the seriousness of such pollution. Moreover, the Thai standard has not been updated to the more stringent international standards: mean levels over a 24-hour limitation and over one year. Therefore, the level of maximum air pollution in Thailand can be 2 to 2.5 times higher than the threshold level, in WHO guidelines [11].

At the end of 2018, more and more Thais became aware of the problem, as they were able to see the impacts of air pollution in Thailand, that was much higher than the recommended threshold: their cities were surrounded by a thick haze frequently and extensively. Bangkok, the capital city of Thailand, was especially affected. It was the most densely populated city, housing 5.7 million people (8.58% of the population) in 2019. Therefore, most, if not all, people living in Bangkok became well aware of the seriousness of air pollution, thus efforts to reduce pollution continued into 2020 and shows no sign of stopping.

We developed a CVD prediction model for Thai population in Bangkok. The model included several machine learning algorithms and was tested to find which algorithm best predicted outcomes. A disease occurrence prediction can be based on a prediction factor, e.g. blood test results. We also added some variables, related to air pollution, to improve the prediction of our model.
II. RELATED WORK

Cardiovascular disease has been studied extensively. Yang et al. found that the existing CVD Risk Prediction Model [3] was specific to one population [6]. However, a CVD prediction model, based on a large population in eastern China, had not yet been developed. At the same time, many existing CVD prediction models were built on the base model, a multi-variable regression model. Therefore, they incorporated complex algorithms, e.g., random forests, to find complex interactions between factors and improve prediction. Their models were based on the culture, lifestyle, behaviour and genetic background of the population in Eastern China. The data was collected from September 2014 to December 2016, from the electronic health record system, tracking 25,231 participants, from a population of 101,056 in Zhejian Province. They used univariate logistic regression and medical expert opinions. Thirty factors were investigated, including gender, age, family income, smoking, excessive drinking, obesity, large waistline and abnormal cholesterol. This group of factors were then modeled by random forests. The model had an Area under the Receiver Operating Characteristic Curve (AUC) of 0.7871, while the multivariate regression model had a lower AUC of 0.7143, indicating a significant increase in the efficiency of the model, when the additional factors were taken into account.

Vathesatogkit et al. developed a model for assessing the risk of cardiovascular disease and the risk of mortality from coronary heart disease and stroke [5]. The model was developed by studying the influences of cardiovascular risk factors, by monitoring the health records of 9,000 Electricity Generating Authority of Thailand employees, for 30 years. The model had an Area under the Curve (AUC) of 0.7871, while the multivariate regression model had a lower AUC of 0.7143, indicating a significant increase in the efficiency of the model, when the additional factors were taken into account.

The European Heart Network [12] highlighted the seriousness of aerosol and air pollution problems affecting CVD. 80% of people died prematurely with CVD due to air pollution and disease. It was Europe’s most common cause of death, with over 3.9 million annual deaths. PM2.5 increased the risk of death from stroke by 19% and from cardiovascular disease by 13%.

So far, there has been no study predicting CVD incidence in Thailand, even though it has faced severe local pollution for a number of years.

III. METHODOLOGY

In this section, we discuss: 1) the nature, extent and acquisition of health and environment data; 2) the feature extraction; 3) the algorithms used; 4) experiment framework and setting; and 5) metrics for evaluating the performance of our model. AUCs and $F_1$-scores were used to measure performance and SHapley Additive exPlanations (SHAP) was used to explain prediction factors [13].

A. Data Collection

The research data was divided into two main areas: health and environment data.

1) Health Data: Data from 368,919 patients, who were Bangkok residents with Thai nationality, admitted at Bangkok Hospital (13.7476912°N, 100.5840781°E), from 2008 to 2019, was collected. The data were rendered anonymous to fully comply with privacy constraints [14]. In total, the data set had 4,751,342 records, which recorded each patient admission. It had three main parts: a) medical history, including admission date, house address, birth date, gender, CVD diagnosis date (year); b) eight vital sign attributes: weight, diastolic, body mass index, height, temperature, systolic, respiration, and pulse; c) other types of clinical data (a total of 2,664 types): e.g., cholesterol, high-density lipoprotein cholesterol, red blood cell, lymphocyte etc.

2) Pollution Data: Environment data were collected from air quality measurement stations of the Thai Pollution Control Department throughout Thailand from 2011–2019 [15]. The data were collected hourly. They consisted of the count of small particulate matter, $\leq 2.5\mu g/m^3$, PM2.5, and the count of particles up to $10\mu g/m^3$, PM10, as well as the levels of CO, O$_3$, NO$_2$ and SO$_2$. The graph in Fig. 1 shows that, for quite some time, the PM2.5 and PM10 counts in Bangkok exceeded levels recommended by the WHO Guidelines [11] and the US EPA [16]. Further, they exceeded the safe levels set by the Thai Health Organization [15] in 2010.

B. Feature Extraction

This section discusses feature extraction used in this study and the CVD risk predictors.

1) Study of Factors Affecting the Disease: Currently, there were many factors or types of clinical data to be considered for each population of patients. However, only a few factors affected CVD risk, e.g., blood pressure and cholesterol. Moreover, patient information in the database was often incomplete. Many values were missing, because each time a patient came to the hospital, they were likely to take tests, different from the ones taken in their last visit, resulting in records, that contained
the clinical data for only one particular test. We then filtered only the records with useful data (types of clinical data). For the first step, we selected only the records with clinical test results. From 4,751,342 records, in the database, only 512,644 had clinical results: 127,667 records of physician-diagnosed CVD patients (positive class) and 384,977 for non-CVD patients (negative class). Type of clinical data contains various of blood tests for various conditions and physician’s written comments features. We set the fraction of missing values (the null value for each feature) to be less than 70% for each class, obtaining a total of 41 features of positive class and 47 features negative class. Then we took features from positive and negative, union together gives all 50 features available, then removed all null values, so that 30,901 records remained.

2) CVD Risk Prediction: To prepare data for predicting CVD risk for the next year, we selected specific patients with four years of retrospective data from 4,751,342 records in the database. There were only 290,476 records that had four years of retrospective: 16,728 records of physician-diagnosed CVD patients (positive class) and 263,748 for non-CVD patients (negative class). Then we created the input by calculating the means (over the preceding year) of the eight vital signs for each patient’s residence in that year. We chose 6-km radius (113.14 km²) because it covers most of the area that we consider in this study. Only Bang Khun Thian district (120.687 km²) that the chosen radius can cover only 94% of the district. We divided the patient into two categories: positive and negative. If the patient did not show the disease in the first to the third year, but contracted in the fourth year, the patient will be classified as positive, otherwise, the patient was classified as negative. At this point, we had four years of data, from years $t - 2$ to $t + 1$. We checked the known result in year $t + 1$ with our prediction from years $t - 2$ to $t$. Thus a sample of our input is shown in Figure 2, with input value from years $t - 2$ to $t$, shown in columns BMI$_{t-2}$ to BMI$_t$ and the prediction for year $t + 1$ in the final column.

C. Models

1) Logistic Regression: Logistic regression is a statistical model that, in its basic form, uses a sigmoid function to model a binary dependent variable: Logistic regression is a low complexity, easily understood, algorithm, so we chose it as the baseline for our experiments.

2) Decision Tree: This is a rule-based algorithm, that repeatedly splits the data set, according to a criterion that maximizes the data separation, resulting in a tree-like structure. The most common criterion is information gain, so that each split minimizes a cost function, e.g. entropy. The major disadvantage of the decision tree is that too many splits of the data set can cause over-fitting and decision trees are sensitive to noisy data. However, compared with the other machine learning methods, a decision tree has the advantage that it can easily be expressed as rules. In many application domains, this advantage outweighs its drawbacks, so that decision trees are widely used in medicine.

3) Random Forest: The random forest is an ‘ensemble learning’ technique, which joins a large number of decision trees, resulting in a reduction of variance, compared to a single decision tree. Random forests have the advantage that they handle noisy and unbalanced data and data can be used without any data reduction.

4) Gradient Boosting Decision Tree: Boosting is another “ensemble learning” technique, that uses multiple classifiers, which it learns sequentially. In each iteration, a Gradient Boosting Decision Tree (GBDT) learns the decision trees by fitting the negative gradients. The main cost in GBDT is the computation time to learn multiple decision trees. We used two variants—XGBoost [17] and LightGBM [18], which were optimized to achieve better performance and reduce computation time.

D. Experiment Framework and Setting

The experiment focused on two main areas: (1) study of factors affecting the disease and (2) predicting risk of disease in the following year.

1) Study of Factors Affecting CVD: We used predictions from XGBoost to determine which features most influenced obtaining correct predictions. Specifically, we selected features to be classified according to their SHAP value.

2) CVD Risk Prediction: CVD predictions were based on two different sets of features: (1) predictions based only on the health features and (2) prediction based on the health and pollution, to improve the prediction.

Even after filtering (see Section III-B2), some data was still missing from the health data, so we removed the features that were missing in more than 20% of the records. The missing data was then populated with the $k$-nearest neighbor algorithm, which compared other similar data and filled in the missing data with $k = 5$. After this, we had 1,711 people remaining, but some patients were in areas not covered by the air quality measuring stations. We, therefore, removed patients for whom we did not have pollution data, leaving us with 1,318 records, divided into 112 people with the disease (positive class) and 1,206 without (negative class).

In this work, 80% of the data set was split into a training set and 20% into a test set. Here, we ran the experiment 10 times with different random splits for training and test sets in both experiments. In each experiment, we used 5-fold cross-validation to obtain a set of optimal parameters with grid search technique for each algorithm. The optimal set of parameter was used to train the optimal model with the training set. Then, the model was evaluated by the test set.
E. Evaluation Metrics

1) $F_1$-score: The $F_1$-score is a way of combining the precision in Eq. (2) and recall in Eq. (3) of the model; it is defined as the harmonic mean of the model’s precision and recall:

$$F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

$$\text{precision} = \frac{TP}{TP + FP},$$

$$\text{recall} = \frac{TP}{TP + FN},$$

where $TP$ is an outcome where the model correctly predicted the positive class, $TN$ is an outcome where the model correctly predicted the negative class, $FP$ is an outcome where the model incorrectly predicted the positive class, and $FN$ is an outcome where the model incorrectly predicted the negative class.

2) Area under the Receiver Operating Characteristic Curve (AUC): An Receiver Operating Characteristic (ROC) curve plots True Positive Rate $TPR$ in Eq. (4) vs False Positive Rate $FPR$ in Eq. (5) at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both $FP$ and $TP$. That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from $(0, 0)$ to $(1, 1)$.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

3) SHAP Value: In applying machine learning algorithms in medicine, it is important that the algorithms is explainable. Algorithms that produce highly accurate, but unexplained, results lead to lack of credibility in the algorithm, even when the results are effective.

Most of the algorithms that can be explained are of low complexity. Thus complex algorithms, that are difficult to explain, may have low credibility. However, using SHAP values can help explain tree based algorithms, for example, decision tree, random forest and GBDT, making the results credible and effective.

IV. Result and Discussion

A. Study of Factors Affecting CVD

We identified people having CVD, with the XGBoost algorithm, using 50 features that were selected (see Section III-B1) to study the factors contributing to CVD. The overall results are ROC curves (see Fig. 3) with mean AUC $0.89 \pm 0.01$—averaged from 10 runs with different random splits.

To determine significant clinical features, we considered the mean absolute SHAP values of the features. Then, the dimension of the data was reduced so that the model was not too complex. In this work, we selected only features that have their mean absolute SHAP value not less than 100 times the maximum mean absolute SHAP value. There were 45 features including Patient Age, Estimated Average Glucose, Cholesterol, Hba1c, Glucose (Fasting), eGFR (African-American), Triglyceride, Creatinine, Monocytes, Uric Acid, MCV, Lymphocytes, HDL-C, LDL-C and others with a mean absolute SHAP value less than the threshold (100 times the maximum SHAP value); these features were used to predict CVD risk next. Fig. 4 shows SHAP values (averaged across 10 runs) of the features. The factor with the highest mean absolute SHAP value is patient age. The blood vessels deteriorate with age and result in an increased risk of CVD. Old adults are at higher risk than young people. Also, CVDs link to diseases, such as diabetes, kidney disease, high blood pressure [19–21]. Blood glucose tests features—hemoglobin A1c (Hba1c), glucose (fasting), and estimated average glucose—are also listed in the top, because chronic high blood sugar levels can cause diabetes and lead to degeneration of small and large blood vessels. In addition, Lipid profile such as high-density lipoprotein cholesterol (HDLC), low-density lipoprotein cholesterol (LDLC) and Triglyceride are significant factors. It causes endothelial dysfunction and can lead to coronary artery disease. Renal function tests with the level of creatinine, triglycerides, uric acid, and eGFR can tell the risk of kidney disease, leading to hypertension disease.

B. CVD Risk Prediction

To predict CVD risk for the next year, we used clinical features (see Section III-D1), there were 37 features, including eight vital sign features and pollution data. Table I shows that, using only health data, led to mean AUC $0.89 \pm 0.03$ and $F_1$-score $0.88 \pm 0.04$, but adding pollution data led to mean AUC $0.91 \pm 0.03$ and $F_1$-score $0.90 \pm 0.03$. Thus prediction with pollution data led to an increase of at least 1% for both mean AUC and $F_1$-score in experiments with every algorithm. With health and pollution data, the most effective algorithm was XGBoost, with mean AUC $0.94 \pm 0.03$ and mean $F_1$-score $0.93 \pm 0.02$. Fig. 5 compares the ROC curve achieved by XGBoost with health features alone and a combination of health and pollution features.

Fig. 6 shows that most features with high mean absolute SHAP values correspond to the results from the study of factors affecting CVD in Section IV-A, such as age, glucose, eGFR, cholesterol, monocytes, creatinine, and features from...
there were 1,318 examples, 299 examples in the respiratory
best machine learning techniques with only health features against
TABLE I: A Comparison of CVD risk prediction by different
algorithm

<table>
<thead>
<tr>
<th></th>
<th>Health Data</th>
<th>Health &amp; Pollution Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
<td>F₁-score</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.86 ± 0.02</td>
<td>0.85 ± 0.01</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.86 ± 0.03</td>
<td>0.82 ± 0.04</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.91 ± 0.02</td>
<td>0.93 ± 0.01</td>
</tr>
<tr>
<td>LightGBM</td>
<td>0.92 ± 0.02</td>
<td>0.93 ± 0.01</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.91 ± 0.02</td>
<td>0.91 ± 0.02</td>
</tr>
<tr>
<td>Average</td>
<td>0.89 ± 0.03</td>
<td>0.88 ± 0.04</td>
</tr>
</tbody>
</table>

vital sign data—systolic, BMI and respiration. According to
the result, respiration was important, with factors for years
(l, t − 2, t − 1) all showing high SHAP values. In the dataset,
there were 1,318 examples, 299 examples in the respiratory
disease group (ICD-10 J00-J99 [22]). In the negative class with
1,206 examples, there were 224 with respiratory disease, and
102 in the positive class, there were 75 respiratory disease
cases, or 71% of the positive class, that is, people with
respiratory disease had a higher rate of respiration than normal
people, causing exposed more pollution than normal people.
Therefore, the respiratory rate has a high SHAP value. In
addition, pollution, e.g. NO₂, PM2.5 were found in the top 50 highest mean absolute SHAP value, that help
predicting the risk of CVD.

V. CONCLUSION

Overall, we found that predicting CVD risk improved with
pollution data, because pollution caused CVD. This concludes
that air pollution problems can affect CVD. Further, five
different machine learning algorithms led to essentially the
same result. However, our scope covered only the population
of Bangkok. In the future, we will extend this work with
patient data from the Bangkok Hospital network and air
quality measurement stations scattered throughout Thailand,
to confirm that our results were not specific to a very large
city, i.e. can be used nationwide.

REFERENCES

[Online], Available: https://www.who.int/en/news-room/
fact-sheets/detail/cardiovascular-diseases-(cvids)
[2] Strategy and Planning Division, Ministry of Public
“The Framingham Heart Study and the epidemiology of
vascular disease: A historical perspective,” Lancet,
2014.
[4] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady,
“The Framingham Heart Study and the epidemiology of
vascular disease: A historical perspective,” Lancet,


